DIGITAL TRANSFORMATIONS FOR HEALTH LAB

GOVERNING HEALTH FUTURES 2030

WORKING PAPER

Apply new AI governance regimes to health first to test their benefits

January 2024

Authors
Anurag Agrawal[1,2], Rohinton Medhora[2,3]

As artificial intelligence (AI) advances, governance systems struggle to catch up. The way countries deal with AI varies, but fundamentally the aim is to strike a balance between potential gains and losses. It is normal for countries and societies to place varying importance on different facets of AI. These include promotion of the innovation sector and its economic benefits; the application of the technology in various sectors; its implication for personal and national security; the privacy and human rights of individuals; misuse, deliberate and unwitting; and ultimately, at the existential level, singularity – the (so far) hypothetical point where the technology becomes uncontrollable and irreversible.

Different approaches to AI governance are emerging around the world

China is the most advanced in rolling out a set of laws and regulations governing AI. Starting with its regulation on recommendation algorithms which came into force in March 2022, the country has since unveiled rules for the management of deep synthesis and generative AI. The public consultation phase for trial measures on the ethical review of advances in science and technology ended in May 2023.

Unlike legislation in Western countries, China's legislative framework is built around *type* of AI technology rather than on the *risk* that crosscuts all AI no matter the specific technology stream that is utilized. Public security concerns and the role of the state are fundamental and underlie all the regulations, and each has specific language to safeguard against discrimination and the spread of negative information to "respect social morality and ethics, abide by business and professional ethics, and follow the principles of impartiality, fairness, openness and transparency, scientific rationality, and honesty", in the words of Article 4 of the algorithm recommendation regulation.

The European Union's AI Act finalized in December 2023 has been termed "visionary" and a "world's first" not because it is literally either, but because it aims to cover all aspects of AI in a single piece of legislation and also serve as a template for a global model of governance for the technology. Rather than be organized along the specific technologies that comprise AI, the legislation classifies risks (always used in the plural) into four categories: minimal or no risks; limited risks; high risks; and unacceptable risks. The first two categories of risks require either no or light regulation with an emphasis on user education and empowerment. Activities deemed high risk carry more obligations and stringent regulations to operate in the EU. The final category, unacceptable risks, contains activities that are banned outright sometimes with limited exceptions. It includes cognitive manipulation, predictive policing, emotion recognition in workplaces and schools, social scoring, and some remote biometric identification systems.

The system is overseen by a European Artificial Intelligence Board, a scientific panel and an advisory forum which together will bring a measure of independence embedded in an official legislative process with appropriate representation of key stakeholders from within the industry and outside it. While much will only be known when the Act is interpreted and applied in individual member countries—and

1 Trivedi School of Biosciences, Ashoka University, Sonepat, Haryana, India.
2 Digital Transformations for Health Lab, University of Geneva, Geneva, Switzerland.
3 Institute for the Study of International Development, McGill University, Montreal, Canada.

there are already critiques that the Act is outdated, or at least not resilient enough to keep up with the pace of advance in the AI field—it, along with China's suite of regulations, is the state-of-the-art in public policy responses to a new and rapidly evolving technological frontier. Other countries are either playing catch up or at this point choosing to not explicitly regulate AI.

Like the EU's approach, the **United States'** approach is also risk-based but varies by sector and is distributed across several federal agencies, with no overarching ethos around managing risks or the industry. Documents like Executive Order 13859 and the AI Bill of Rights issued by the White House Office of Science and Technology Policy provide a framework, reiterate risk-based and sector-based governance, but implementation is fragmented across the federal government and is sometimes no more than aspirational.

Canada's proposed Artificial Intelligence and Data Act (AIDA) is also patterned on the EU's risk-based approach with escalating penalties for non-compliance but has run into a flurry of criticism centred on three issues. First, most major concepts (such as "high impact systems" and "material harm") are undefined and will be defined later via administrative decree thus circumventing Parliamentary scrutiny during debate of the proposed legislation. Second, government departments and entities are exempt from its provisions. Third, all power is vested in the Minister of Industry and departmental officials from a Ministry historically concerned with overseeing industrial policy rather than broader questions of the public good, privacy and human rights. A series of amendments to the original proposal has just been tabled but the core issues in play illustrate how complicated the matter is, and what dilemmas and trade-offs all countries face when it comes to managing AI.

The US approach provides for more flexibility (across sectors). A presidential executive order signed in October 2023 requires that safety test results for AI-based applications in healthcare be shared with the government, and that methods be developed to protect against the risks of using AI to engineer dangerous biological materials. The new directive also calls for voluntary commitments from companies working in the field of AI to develop methods that ensure the safety and personal security of US citizens.

The EU's and China's comprehensive approach are likely to encourage certainty, consistency and stability in governance, thus encouraging trust and investment in the AI field. On balance the EU puts a premium on precaution, favouring regulation over promoting innovation while the Chinese and US approaches tilt towards the other side.

In most **other parts of the world,** AI governance tends to be guidance and aspirational and is at early stages of anything beyond this. In August 2023, the **UK** government presented to Parliament a Policy Paper that proposes a risk-based, "pro-innovation approach to AI regulation" that is likely to be a light touch on most of the spectrum of AI use, and is months away from implementation. In April 2023, **India**'s Ministry of Electronics and IT said that it does not intend to introduce legislation to regulate the growth of AI but will implement necessary policies and infrastructure to cultivate a robust AI sector in the country. In addition to a set of Principles for Responsible AI enunciated in 2021, a Digital India Act (forthcoming) will replace the Information Technology Act of 2000. Its core constituents will be online safety, trust and accountability, open internet, and regulations of new age technologies like AI and blockchain technologies. In relation to AI, the legislation may delineate specific "no-go areas" for companies and internet intermediaries employing AI and machine learning in consumer-facing applications with penalties for non-compliance.

Such approaches may be seen as a plus, particularly for countries intending to nurture a high-tech innovation sector or those with limited capacity to develop policies and enforce them—if regulations are not baked in, there is room to leave spaces for innovation and flexibility.

Health: a sector where the consequences of AI governance are great

The overall picture of AI governance is one of experimentation, significant gaps in coverage, and with it a risk of either over- or under-regulation of a fast-moving game-changing *general purpose* technology. But AI's impacts, good and bad, will vary across sectors, geographies and time periods. Existing attempts to regulate AI are thus a platform on which to build out, dealing in specific cases and situations where the technology is applied.

The health sector is a good example of one where the potential gains from the proper use of Big Data and AI are high, as are the potential risks to personal and public safety, privacy and human rights. The demand for health services outstrips available supply by far, leading to great expectations from AI by the public, providers and policymakers. However, the complexity and contextuality of patient needs makes current AI systems only suitable for use under human supervision, despite large recent strides in generative and foundational AI. The few examples where regulators have permitted autonomy to AI systems have been in very specific contexts such as normal chest X-rays where it is clear that AI algorithms are better than humans at picking abnormalities. Abnormalities are still referred to human overseers for final determination. Despite hype about self-learning AI, which would improve with use, current regulations explicitly prohibit any change to AI algorithms after approval, which is typically bounded to specific use cases. Portability of health-AI solutions to populations other than those on whom they were trained has also been suboptimal and thus the onus of their use typically vests with the physician, which is an unsatisfactory governance system equivalent of passing the buck. Regulatory mechanisms paving the way for approving autonomous AI systems that can truly overcome the shortage of healthcare personnel and improve on human-based diagnostic outcomes—even in cases where there isn't a shortage of qualified personnel—require two distinct spheres of improvement. First, in AI itself, with strong foundational models that have generative capacity but are anchored in relevant knowledge and are free from hallucinations. Second, in the data that goes into the foundational knowledge.

Presence of diverse datasets of suitable quality, with equitable representation of different patient groups, is a necessary step in the development and testing of future AI solutions. However, this is far from reality and the problem is even more severe when looking at available data for non-white populations. Existing science is often based on clinical trials using datasets that are not representative of the diversity in populations, so AI systems based on such studies bake in current biases in our knowledge. There are three distinct aspects to this: digital data generation, standardized data formats, and availability of usable data. Rapidly expanding digital transformation and a concerted move towards common data models and dictionaries is likely to increase usable digital data, but there are still severe governance challenges in making such data available. Unresolved ownership questions are a major contributor to this data lethargy, as is data-protectionism for economic or strategic interests.

The ownership of health data varies across the globe. While US federal law gives patients legal privacy, security and accuracy rights related to their health data, they are not treated as owners. Personal data laws vary across states and consumers typically do not have rights to demand that their data residing with service providers not be used without their permission. Current interpretations typically allow service providers the right to use such data to improve services, which can have a wide meaning in health. In contrast, the EU is more explicit about ownership vesting with the patient, who has the right to restrict its usage by others, including after deidentification. Other countries, such as India, place importance on

health data privacy and acknowledge ownership, but have permissive laws for use of deidentified data. Given that true deidentification is nearly impossible, especially for data like genomes, this may not be sufficient. There have been attempts to bypass this problem altogether by using purpose-limitation and data solidarity, rather than ownership and consent, as the ethical underpinning, alongside technical solutions like federated learning that allow AI algorithm development without any direct access to discrete identifiable data.

Broad approaches to AI governance should be tested against universally agreed goals

To conclude, current approaches to AI regulation are inadequate in two dimensions. On the one hand, cognizant of AI's "general purpose" nature, actual and forthcoming legislation is also general, in the sense that it does not distinguish between the various uses or sectors to which AI is applied. On the other hand, the approaches vary in intent, scope and coverage; we are far from a universal approach to AI that facilitates cross-country cooperation. In its interim report issued in December 2023, the UN's AI Advisory Board sets out some guiding principles that provide a frame for interoperability in the use of AI for the public good.

The challenge is to translate broad intention into a viable way forward. This can only be done by testing a governance regime in a single sector. The health sector is a good place to start. If we started with a clear goal that would be near-universally accepted—say improved (i.e. faster, more accessible, more accurate) outcomes in imaging, or a deep understanding of all known protein structures—and worked backwards on the data governance and technology management requirements for this to be achieved, then this would put flesh on the bones of current initiatives to manage AI, while also providing a way forward in other areas of application. A good starting point would be the creation of a so-called data trust, a mechanism to marshal data for an explicitly agreed purpose, defined governance and shared benefits among the parties constituting the agreement.

While daunting, the challenge is not entirely unique. Similar, though of course not exactly the same, hurdles were faced by countries (or more accurately scientific and policy communities in them) in dealing with biotechnologies, particle acceleration and exploring outer space, to name three. In each case an openness to cooperation, flexibility and experimentation within bounded limits has resulted in multi-country endeavours that have provided significant benefits, direct and spun off, to humankind. Done right, a similar approach to the application of AI in health will advance the larger agenda of stimulating the roll out of a technology that not only promises transformational good but is seen by all concerned to do just that.

Acknowledgments: Thanks to Louise Holly for her support with background research.

Suggested citation: Agrawal A, and Medhora R. (2024) Working paper: Apply new AI governance regimes to health to first test their benefits. Geneva: Digital Transformations for Health Lab

Digital Transformations for Health Lab (DTH Lab)

Hosted by: The University of Geneva
Campus Biotech- Chemin des Mines 9,
1202 Geneva, Switzerland

www.governinghealthfutures2030.org